Historic, archived document

Do not assume content reflects current scientific knowledge, policies, or practices.

A99.9
F764Un
Cop. 2

Research Note

UNITED STATES DEPARTMENT OF AGRICULTURE
FOREST SERVICE
INTERMOUNTAIN FOREST & RANGE EXPERIMENT STATION
OGDEN, UTAH 84401

U.S.D.A. Forest Service
Research Note INT-79
1968

COMPUTER PROGRAMS FOR WEIGHTING LENGTH-OF-STAY AND OTHER VISITOR CHARACTERISTICS

J. Alan Wagar and Joel F. Thalheimer[1]

U. S. DEPT. OF AGRICULTURE
NATIONAL AGRICULTURAL LIBRARY
MAY 28 1970
CURRENT SERIAL RECORDS

ABSTRACT

When recreationists of different characteristics have different probabilities of being included in recreation surveys, the resulting averages for these characteristics can be severely biased unless special computational procedures are used. This paper describes the use of two computer programs for correcting biases that would otherwise result from (a) differences in length of stay, (b) known probabilities for sampling each individual, or (c) the combined effects of a and b.

If visitors of different ages, incomes, education levels, and other characteristics have different probabilities of being sampled in a recreation survey, results can give strongly biased averages unless special computational procedures are used. Both length of stay and sampling design can affect the probability of including each visitor in the sample.

As Lucas pointed out, visitors who stay a long time are much more likely to be sampled than those who stay only a short time (3, 4).[2] As a result, the characteristics associated with long stays are overrepresented in the data and characteristics associated with short stays are underrepresented. Also, if different recreation sites are sampled at different intensities, the characteristics of visitors on the more heavily sampled sites would be overrepresented in the data. However, Lucas showed that averages for visitor characteristics can be computed without bias from survey data if each observation is weighted by the inverse of its probability of being included in the sample.

---

[1] The senior author formerly was Leader of the Cooperative Recreation Research project maintained by the Intermountain Forest and Range Exp. Station, Forest Service, U.S. Department of Agriculture, in cooperation with Utah State University at Logan; and the junior author was a graduate student in the Department of Forest Science at the same University. Wagar is now Leader of the Cooperative Recreation Research project maintained by the Pacific Northwest Forest and Range Exp. Station in cooperation with the University of Washington at Seattle; and the junior author is now owner-manager of a resort in Arizona.

[2] Numbers in parentheses refer to Bibliography.

The extent of possible bias is shown by a study of fisherman use at Flaming Gorge Reservoir in Utah and Wyoming (2). Unweighted estimates of average length of stay were from 2.60 to 3.44 times as large as estimates that were weighted to correct for the probability of including each person in the sample.

Visitor characteristics related to length of stay could also be overestimated from unweighted survey data. For example, an individual's expenditures usually depend on how long he stays, and average expenditures per person could therefore be grossly overestimated from unweighted survey data.

Similar problems could occur in estimates of age, income, education, and other visitor characteristics. For example, if retired people tended to stay longer on a recreation site than people from other age groups, the average age of visitors could be substantially overestimated from unweighted data.

Deming provided procedures not only for obtaining weighted averages but also for computing the variances for such averages (1). The following formulae are based on these procedures:

$$\text{Weighted mean} = \overline{X}_w = \frac{\Sigma w_i X_i}{\Sigma w_i}$$

$$\text{Variance of the mean} = S^2_{\overline{X}} = \frac{\Sigma\left(w_i X_i^2\right) - \overline{X}_w^2\ \Sigma w_i}{\Sigma w_i\ (n-1)}$$

where:

w_i is the weighting factor for individual i. This is the inverse of the individual's probability of being included in the sample.

X_i is an observation or measurement of a variable for individual i.

n is the number of individuals in the sample.

These weighting procedures have been incorporated into two short computer programs that permit rapid processing of data suited to any of seven different weighting schemes. In these programs, up to 10 variables (characteristics) per sampled individual can be weighted by the inverse of (a) length of stay, (b) a known sampling probability, or (c) the combined effects of (a) and (b). With very slight modification, the programs will readily handle additional variables for each sampled individual. A weighted mean and variance are then computed and printed for each variable (labeled as "datum" in the printed output). The programs also compute and print a weighted mean and variance for length of stay.

PROGRAM 1

For flexibility, six sampling schemes (A to F) are provided in Program 1 and can be selected to suit various sampling situations. Schemes A, B, and C correct only for length of stay.

For Scheme A, the smallest unit of time recognized is a day, and each observation is corrected for the number of days an individual is present. This scheme would be appropriate for an area that is sampled (a) during the hours people normally arrive and depart and (b) on each of a random selection of days. Scheme A requires a minimum amount of detail in the data, and the output includes a weighted mean and variance for the days on which a visitor is actually present. (This is not quite equivalent to length of stay in days but is a close approximation.) Sample results are given in figure 1.

```
TEST RUN FOR PROGRAM 1, SCHEME A OR D

NUMBER OF SAMPLES  =    10.

WEIGHTED AVERAGE, DAYS PRESENT  =     3.1683
VARIANCE FOR LENGTH OF STAY  =      9.5161

WEIGHTED AVERAGE FOR DATUM  1  =   42.4583
WEIGHTED VARIANCE OF THE MEAN FOR DATUM  1  =     756.5978

WEIGHTED AVERAGE FOR DATUM  2  =     2.4979
WEIGHTED VARIANCE OF THE MEAN FOR DATUM   2 =       1.2085

WEIGHTED AVERAGE FOR DATUM  3  =    7.0351
WEIGHTED VARIANCE OF THE MEAN FOR DATUM  3  =       3.0713

WEIGHTED AVERAGE FOR DATUM  4  =   39.1554
WEIGHTED VARIANCE OF THE MEAN FOR DATUM  4  =     127.4493

WEIGHTED AVERAGE FOR DATUM  5  =   42.9808
WEIGHTED VARIANCE OF THE MEAN FOR DATUM  5  =      32.3951
```

Figure 1.--Simulated sample printout for Program 1 showing results as provided by Schemes A and D.

In Scheme B, a sampling period is defined to include the same hours of each sampled day. For example, campers might be sampled between 2000 and 2200 hours in the evening on a random selection of sample days; thus for weighting, each observation is corrected for the number of 2000-to-2200 periods during which an individual is present. To simplify the data and computations, all times are given on a 24-hour basis. Thus 1100 is 11 a.m., 1200 is 12 noon, 1300 is 1 p.m. ... and so forth.

Scheme B also computes the actual number of on-the-hour times during which each sampled individual is present. Thus a person arriving at 1510 and leaving at 2150 would be present for six on-the-hour times--1600 through 2100. As shown in figure 2, the weighted average for hours present is printed out along with its variance.

In Scheme C, sampling takes place on single on-the-hour times selected at random within specified limits. For example, sampling on a recreation site might be specified for each of 20 on-the-hour times selected at random between the hours of 0900 and 2100 on any day between May 30 and September 6. In this scheme, each observation would be weighted to correct for the number of on-the-hour times an individual is present during which sampling could take place. Thus, if he arrived at 1630 one day and left at 0930 the next, he would be present for the six on-the-hour sampling times--1700 through 2100 and 0900. As in Scheme B, the average and variance for hours present are printed in Scheme C, as illustrated in figure 2.

Schemes D, E, and F of Program 1 are designed to handle situations where it is desirable to weight sample data to correct for known sampling probabilities in addition to the probabilities created by length of stay.

```
TEST RUN FOR PROGRAM 1, SCHFME B, C, E, OR F.

NUMRER OF SAMPLES  =    10.

WEIGHTED AVERAGE, HOURS PRESENT  =    11.1536
VARIANCE FOR LENGTH OF STAY  =     862.7324

WEIGHTED AVERAGE FOR DATUM  1  =    12.1405
WEIGHTED VARIANCE OF THE MEAN FOR DATUM   1  =    134.9237

WEIGHTED AVERAGE FOR DATUM  2  =     1.0292
WEIGHTED VARIANCE OF THE MEAN FOR DATUM   2  =      0.3976

WEIGHTED AVERAGE FOR DATUM  3  =     4.6906
WEIGHTED VARIANCE OF THE MEAN FOR DATUM   3  =      1.3208
```

Figure 2.--Simulated sample printout for Program 1 showing results as provided by Schemes B, C, E, and F.

For example, it might be convenient to sample large recreation areas more often than small recreation areas while still correcting for length-of-stay bias. In this case the characteristics of each sampled individual would need to be weighted for both the probability of sampling a given recreation site and the probability of including an individual in the sample at the site selected. Therefore w_i in Deming's formulae becomes the inverse of the product of two sampling probabilities. With minor modification, Program 1 would readily handle additional probabilities. The weighting factor would simply be the inverse of the product of three or more probabilities.

Scheme D is equivalent to Scheme A except that weighting corrects for known sampling probabilities in addition to the number of days in which each sampled individual is present. In the same manner, Scheme E is a modification of Scheme B and Scheme F is a modification of Scheme C. Scheme E weights for known probabilities in addition to the number of sampling periods during which a sampled individual is present. Scheme F weights for known probabilities in addition to the number of on-the-hour times an individual is present during which he could be sampled. Figure 2 also illustrates the form of output resulting from Schemes D, E, and F.

PROGRAM 2

Program 2 is a simplification of Program 1 designed for data in which the probability of each individual's being sampled is already known and can be entered directly with the rest of the sample data. For example, if a survey is made of visitors leaving a recreation area at various road exits, it might be desirable to sample heavily used exits on 10 percent of the days in a use season and lightly used exits on only 5 percent of the days. In this kind of survey, length of stay would not affect sampling probabilities. Thus the probability of sampling each person leaving by a heavily used exit would be 0.10 and the probability of sampling each person leaving by a lightly used exit would be 0.05. Figure 3 shows a sample output for Program 2.

```
TEST RUN FOR PROGRAM 2.

NUMBER OF SAMPLES  =     5.

WEIGHTED AVERAGE FOR SAMPLING PROBABILITIES  =     2.6592
VARIANCE FOR SAMPLING PROBABILITIES  =    11.1292

WEIGHTED AVERAGE FOR DATUM  1  =    11.0123
WEIGHTED VARIANCE OF THE MEAN FOR DATUM   1  =      39.8680

WEIGHTED AVERAGE FOR DATUM  2  =    67.7673
WEIGHTED VARIANCE OF THE MEAN FOR DATUM   2  =     250.5530
```

Figure 3.--Simulated sample printout for Program 2.

PROGRAM CHARACTERISTICS

Programs 1 and 2 were written in FORTRAN IV for a Control Data Corporation 3200 Computer and were then adapted to an IBM 7094.[3] However, they can be adapted to almost any computer with a FORTRAN IV compiler. Because few data are held in memory at any one time, little storage is needed beyond that required for storing the program itself.

Variable names have been used to indicate the card reader (IZIN) and the printer (IZOUT). As a result, the programs will fit many computers immediately if IZIN is set equal to the number designating the card reader and IZOUT is set equal to the number designating printer.

---

[3] Trade names are used herein solely for identification and do not imply endorsement by the U.S.D.A. Forest Service.

INSTRUCTIONS FOR PROGRAM USE

Program 1

The first three data cards of Program 1 read in (a) the days to the beginning of each month, (b) a label for whatever job is being run, and (c) the values of J, B, E, Q, and QW. For all data that do not include the date February 29, the first card will be punched 000031059090...334 in columns 1 to 36, indicating that 000 days in the year have passed at the beginning of January, 031 days have passed at the beginning of February, 059 at the beginning of March, 090 at the beginning of April, etc., to 334 at the beginning of December. For data that include February 29, this card would be replaced by one reading 000031060091...335. Program 1 will handle data that extend from one year to the next, provided that no more than 365 days are included.

The second data card can have any combination of letters and numbers. This is used for a label of up to 80 characters and spaces that identifies the job being run.

The third data card specifies values of J, B, E, Q, and QW in columns 7-8, 9-10, 11-12, 13-14, and 15-16, respectively. Variable J is set equal to the number of variables to be weighted for each sampled individual (from 01 to 10); B specifies the earliest hour that sampling can occur (01 to 24); and E specifies the latest hour that sampling can occur (01 to 24). The variables Q and QW are selected to specify computations appropriate to the various sampling schemes. Thus if QW is set equal to +1, setting the variable Q equal to -1, 00, or +1 will specify computations for Scheme A, B, or C, respectively. If QW is set equal to zero, then setting the variable Q equal to -1, 00, or +1 will specify computations for Scheme D, E, or F, respectively.

Table 1 summarizes the characteristics of sampling Schemes A to F and coding for the third data card used with Program 1.

After the first three data cards, an additional data card is used for each sampled individual. Each of these cards specifies time of arrival, time of departure, and values for up to 10 variables. For Schemes D, E, and F, each card must also specify a known sampling probability (P).

In Schemes B, C, E, and F, arrival and departure times are coded as 6-digit numbers. The first two digits specify month (F1 for arrival, AL1 for departure), the second two digits specify day of month (F2 for arrival, AL2 for departure), and the third two digits specify hour (F3 for arrival, AL3 for departure). The hour of arrival is recorded as the first on-the-hour time a visitor is present, and the hour of departure is recorded as the last on-the-hour time a visitor is present. Thus an arrival time of 1410 would be coded as 15, a departure time of 0945 as 09, etc.

Program statements 3 and 4 can be rewritten to read data in whatever format is desired. For example, the hour of arrival (F3) and the hour of departure (AL3) are not needed for Schemes A and D, and program statements 3 and 4 could be simplified to delete these variables.

For all six sampling schemes in Program 1, a blank card is placed after the last data card. This causes the program to go on to the remaining computations and then stop.

If all data are to be handled according to Scheme A or D, Program 1 can be shortened by deleting all program statements preceded by X and by changing the statement preceding Statement 6 from GO TO 8 to GO TO 34. For this shortened program, the value of J must still be entered in columns 7-8 and the value of QW entered in columns 15-16 of the third data card, but values of B, E, and Q are not used. Further simplification of Program 1 results in Program 2.

Table 1.--Summary of characteristics for Sampling Schemes A to F of Program 1 and of the coding for the third data card used with this program

Sampling scheme	Weights for:	Coding for 3rd data card				
		J	B	E	Q	QW
A	length of stay as days present	number of observations per sampled individual	not used	not used	-1	+1
B	length of stay as sampling periods present	ditto	1st hour on which sampling can take place	last hour on which sampling can take place	00	+1
C	length of stay as on-the-hour sampling times present	ditto	ditto	ditto	+1	+1
D	length of stay (as days) plus known sampling probability	ditto	not used	not used	-1	00
E	length of stay (as sampling periods) plus known sampling probability	ditto	1st hour on which sampling can take place	last hour on which sampling can take place	00	00
F	length of stay (as hours) plus known sampling probability	ditto	ditto	ditto	+1	00

Program 2

Application of Program 2 is similar and somewhat simplified from application of Program 1. No data card is used to specify days to the first of each month so the first data card is the one used to provide a label of up to 80 characters and spaces. Following this, one card is used for each sampled individual to give the sampling probability (P) and up to 10 variables for that individual. As before, a blank card is used at the end of the data to direct the computer to remaining computations and then to a stop.

PROGRAM 1 -- PAGE 1

```
          DIMENSION X(10),SWIZI(10),VARP(10),A(12),ID(20)
          IZIN=5
          IZOUT=6
    C  IZIN=CARD READER.  IZOUT=PRINTER
          READ(IZIN,25) (A(L),L=1,12)
       25 FORMAT(12F3.0)
          READ(IZIN,100)(ID(I),I=1,20)
     100  FORMAT (20A4)
          WRITE(IZOUT,101)(ID(I),I=1,20)
     101  FORMAT(1H1,20A4)
      52  READ(IZIN,32) J,B,E,Q,QW
       32 FORMAT( 6X,I2,4F2.0)
          Z=0.
          SWIZ=0.
          VARPZ=0.
          SWIJ=0.
          DO 5K=1,J
          SWIZI(K)=0.
        5 VARP(K)=0.
     3    READ(IZIN,4)P,F1,F2,F3,AL1,AL2,AL3,(X(I),I=1,J)
     4    FORMAT(6X,F4.3,6F2.0,10F5.2)
          IF(F1+AL1)19,18,19
       19 IF (F1-AL1)23,6,26
       26 AL2=AL2+365.
       23 L=F1
          FF=A(L)+F2
          L=AL1
          FFF=A(L)+AL2
          T=FFF-FF+1.
          GO TO 8
        6 T=AL2-F2+1.
X       8 IF(Q) 34,1,1
X       1 IF(1.+E-F3)9,9,40
X      40 IF(Q) 41,41,10
X      41 AF3=1.+E-B
X         GO TO 11
X     10  IF(B-F3)42,42,43
X      42 AF3=1.+E-F3
X         GO TO 11
X      43 AF3=1.+E-B
X         GO TO 11
X       9 AF3=0.
X      11 IF(AL3-B+1.)12,12,44
X      12 AL3A=0.
X         GO TO 16
X      44 IF(Q) 15,15,13
X     13  IF(AL3-E)14,14,15
X      14 AL3A=AL3-B+1.
X         GO TO 16
X      15 AL3A=E-B+1.
X         GO TO 16
       34 WIJ=1./T
          ZI=T
X         GO TO 35
X      16 WIJ=1./(AF3+AL3A+((T-2.)*(E-B+1.)))
```

PROGRAM I -- PAGE 2

```
               ZI=25.-F3+AL3+((T-2.)*24.)
         35    IF(QW)61,60,61
         60    WIJ=WIJ/P
         61    DO 17 K=1,J
               SWIZI(K)=WIJ*X(K)+SWIZI(K)
            17 VARP(K)=WIJ*(X(K)**2)+VARP(K)
               SWIJ=WIJ+SWIJ
               SWIZ=WIJ*ZI+SWIZ
               VARPZ=WIJ*(ZI**2)+VARPZ
               Z=Z+1.
               GO TO 3
          18   WRITE(IZOUT,21) Z
            21 FORMAT(21HONUMBER OF SAMPLES = F7.0)
               SWIZ=SWIZ/SWIJ
               VARPZ=(VARPZ-((SWIZ**2)*SWIJ))/((Z-1.)*SWIJ)
   X           IF(Q)36,37,37
          36   WRITE(IZOUT,38)SWIZ
            38 FORMAT(35HOWEIGHTED AVERAGE, DAYS PRESENT  = F10.4)
   X           GO TO 39
   X      37   WRITE(IZOUT,31)SWIZ
   X        31 FORMAT(35HOWEIGHTED AVERAGE, HOURS PRESENT = F10.4)
          39   WRITE(IZOUT,33) VARPZ
            33 FORMAT(31H VARIANCE FOR LENGTH OF STAY = F12.4)
               DO2OK=1,J
               SWIZI(K)=SWIZI(K)/SWIJ
               VARP(K)=(VARP(K)-((SWIZI(K)**2)*SWIJ))/((Z-1.)*SWIJ)
               WRITE(IZOUT,22)K,SWIZI(K)
            22 FORMAT(28HOWEIGHTED AVERAGE FOR DATUM I2,3H = F10.4)
          20   WRITE(IZOUT,30)K,VARP(K)
            30 FORMAT(41H WEIGHTED VARIANCE OF THE MEAN FOR DATUM I2,3H = F12.4)
               READ(IZIN,50)INT
          50   FORMAT(I6)
               IF(INT-999999)52,51,52
          51   STOP
               END
```

PROGRAM 2

```
      DIMENSION X(10),SWIZI(10),VARP(10),A(12),ID(20)
      IZIN=5
      IZOUT=6
C  IZIN=CARD READER.  IZOUT=PRINTER
      READ(IZIN,1)(ID(I),I=1,20)
    1 FORMAT(20A4)
      WRITE(IZOUT,2)(ID(I),I=1,20)
    2 FORMAT(1H1,20A4)
   52 READ(IZIN,32)J
   32 FORMAT(10X,I2)
      Z=0
      SWIZ=0
      VARPZ=0
      SWIJ=0
      DO 5 K=1,J
      SWIZI(K)=0
    5 VARP(K)=0
    3 READ(IZIN,4)P,(X(I),I=1,J)
    4 FORMAT(10X,F6.3,10F6.2)
      IF(P)34,37,34
   34 WIJ=1./P
      ZI=P
      DO 17 K=1,J
      SWIZI(K)=WIJ*X(K)+SWIZI(K)
   17 VARP(K)=WIJ*(X(K)**2)+VARP(K)
      SWIJ=WIJ+SWIJ
      Z=Z+1.
      VARPZ=WIJ*(ZI**2)+VARPZ
      GO TO 3
   37 WRITE(IZOUT,40)Z
   40 FORMAT(21HONUMBER OF SAMPLES = ,F7.0)
      SWIZ=Z/SWIJ
      VARPZ=(VARPZ-((SWIZ**2)*SWIJ))/((Z-1.)*SWIJ)
      WRITE(IZOUT,38)SWIZ
   38 FORMAT(46HOWEIGHTED AVERAGE FOR SAMPLING PROBABILITIES =,F10.4)
      WRITE(IZOUT,33)VARPZ
   33 FORMAT(38H VARIANCE FOR SAMPLING PROBABILITIES =,F12.4)
      DO 20 K=1,J
      SWIZI(K)=SWIZI(K)/SWIJ
      VARP(K)=(VARP(K)-((SWIZI(K)**2)*SWIJ))/((Z-1.)*SWIJ)
      WRITE(IZOUT,22)K,SWIZI(K)
   22 FORMAT(28HOWEIGHTED AVERAGE FOR DATUM I2,3H = F10.4)
   20 WRITE(IZOUT,30)K,VARP(K)
   30 FORMAT(41H WEIGHTED VARIANCE OF THE MEAN FOR DATUM I2,3H = F12.4)
      READ(IZIN,50)INT
   50 FORMAT(I6)
      IF(INT-999999)52,51,52
   51 STOP
      END
```

VARIABLE NAMES FOR PROGRAMS 1 AND 2

Variable	Meaning
A(L)	Days in year that have elapsed up to the beginning of month L.
ID(I)	Provision for an alphanumeric label of up to 80 characters.
J	Number of different observations or measurements made for each individual in the sample.
B	The first on-the-hour time that sampling can begin. This is coded from 01 to 24.
E	The last on-the-hour time that sampling can occur. This is also coded from 01 to 24.
Q and QW	Values coded as -1, 00, or +1 and 00 or +1, respectively, that permit computations to be directed to Scheme A, B, C, D, E, or F.
Z	A counter to record number of individuals in sample.
SWIZ	Summation of weighting factor times length of stay (in Program 1) or other sampling probability (Program 2). SWIZ also is used as this summation divided by the summation of the weighting factors.
VARPZ	Summation of weighting factors times squared lengths of stay (in Program 1) or times squared sampling probabilities (Program 2). VARPZ is also used as the variance for lengths of stay (Program 1) and as variance for sampling probabilities (Program 2).
SWIJ	Summation of the weighting factors.
SWIZI(K)	Summation of weighting factors times values of a variable. SWIZI(K) is also used as this summation divided by SWIJ.
VARP(K)	Summation of weighting factors times squared values of a variable. VARP(K) is also used as variance for such an observation or measurement.
F1	Month of arrival.
F2	Day of month of arrival.
F3	Hour of day of arrival.
AL1	Month of departure.
AL2	Day of month of departure.
AL3	Hour of day of departure.
X(I)	Value of variable I as observed for each individual in the sample. (I goes from 1 to j.)
AF3	Number of hours of possible exposure to sampling on day of arrival.
AL3A	Number of hours of possible exposure to sampling on day of departure.
T	Number of days on which present.
WIJ	The weighting factor. This equals the inverse of the sampling probability.
ZI	The probability of being included in the sample.
P	The probability of being included in the sample.
INT	A variable used at end of data to stop program.

BIBLIOGRAPHY

1. Deming, W. Edwards.
   1944. Statistical Adjustment of Data, pp. 18-20. New York: John Wiley & Sons.
2. Hewston, John G.
   1965; 1966. Unpublished reports on file at Utah Cooperative Fishery Unit, Utah State Univ., Logan, Utah.
3. Lucas, Robert C.
   1963. Bias in estimating recreationists' length of stay from sample interviews. J. Forest. 61:912-914.
4. ________, and Dennis L. Schweitzer.
   1965. Outdoor recreation surveys: length-of-stay bias and its correction by computer. U.S. Forest Serv. Res. Note LS-68, 2 pp.

The Forest Service of the U. S. Department of Agriculture is dedicated to the principle of multiple use management of the Nation's forest resources for sustained yields of wood, water, forage, wildlife, and recreation. Through forestry research, cooperation with the States and private forest owners, and management of the National Forests and National Grasslands, it strives — as directed by Congress — to provide increasingly greater service to a growing nation.